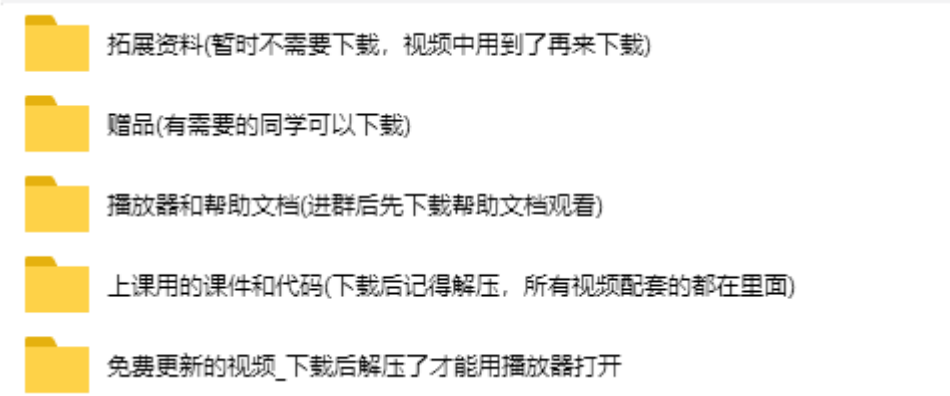
第九讲:分类模型

本讲将介绍分类模型。对于二分类模型，我们将介绍逻辑回归(logistic regression)和Fisher线性判别分析两种分类算法；对于多分类模型，我们将简单介绍Spss中的多分类线性判别分析和多分类逻辑回归的操作步骤。

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

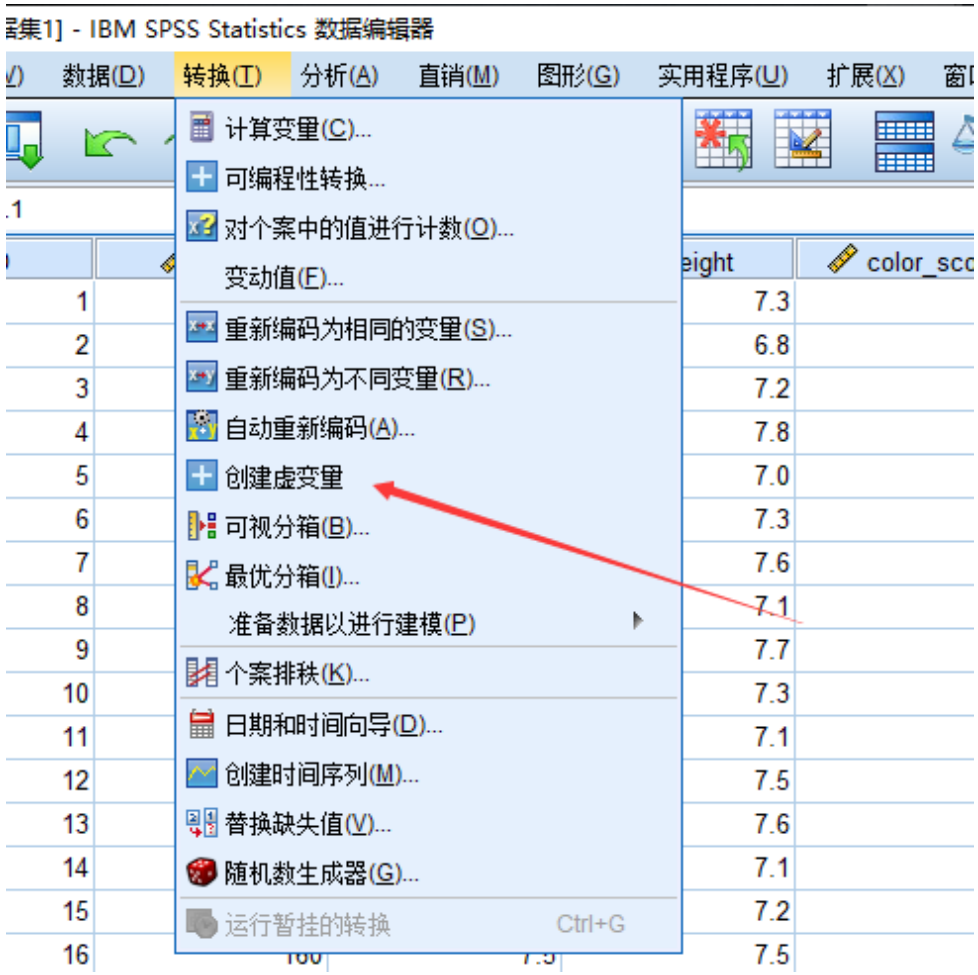
水果分类的例子

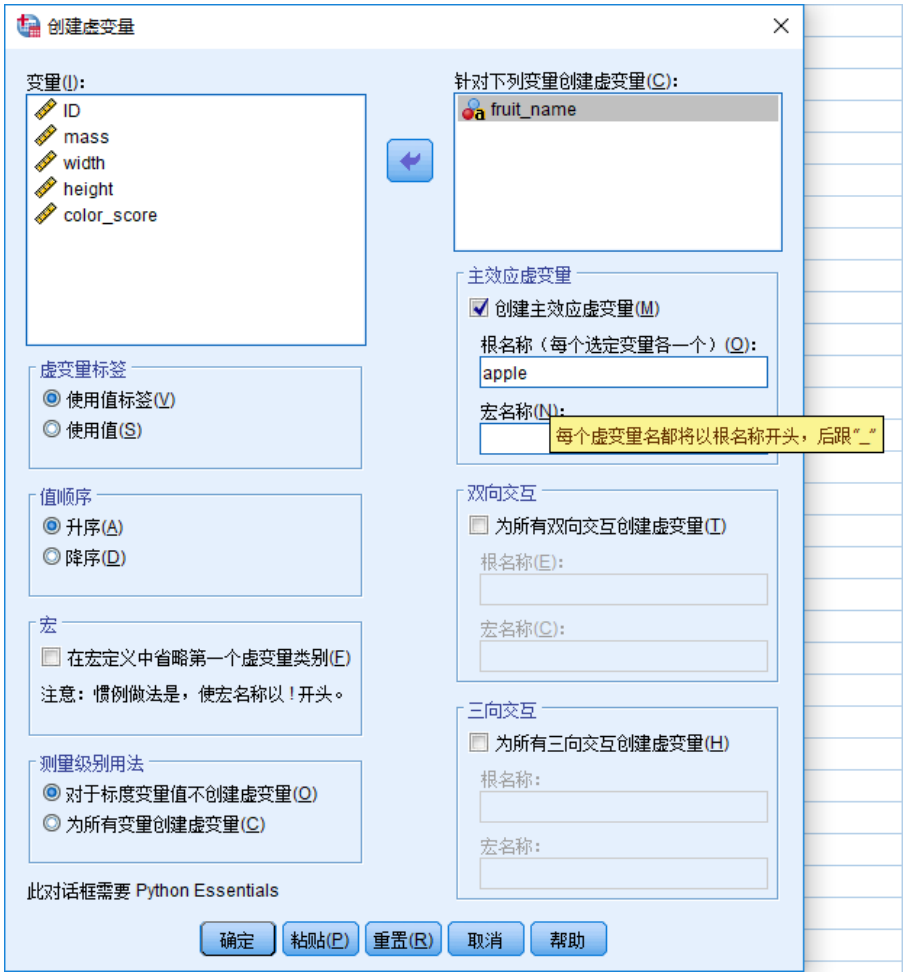
根据水果的属性，判断该水果的种类。

mass: 水果重量
width: 水果的宽度
height: 水果的高度
color_score: 水果的颜色数值，范围0-1
fruit_name：水果类别

前19个样本是苹果
后19个样本是橙子
用这38个样本预测后四个样本对应的水果种类。

数据预处理：生成虚拟变量





逻辑回归logistic regression

类型	模型	Y的特点	例子
线性回归	OLS、GLS（最小二乘）	连续数值型变量	GDP、产量、收入
0-1回归	logistic回归	二值变量（0-1）	是否违约、是否得病
定序回归	probit定序回归	定序变量	等级评定（优良差）
计数回归	泊松回归（泊松分布）	计数变量	每分钟车流量
生存回归	Cox等比例风险回归	生存变量（截断数据）	企业、产品的寿命

对于因变量为分类变量的情况，我们可以使用逻辑回归进行处理。
把y看成事件发生的概率，$y \geq 0.5$表示发生；$y < 0.5$表示不发生

线性概率模型

线性概率模型（Linear Probability Model，简记LPM）

直接用原来的回归模型进行回归。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

写成向量乘积形式：$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + u_i \ (i = 1, 2, \cdots, n)$

内生性问题：y_i只能取0或者1（回归系数估计出来不一致且有偏）

$$u_i = \begin{cases} 1 - \boldsymbol{x_i'}\boldsymbol{\beta} & , y_i = 1 \\ -\boldsymbol{x_i'}\boldsymbol{\beta} & , y_i = 0 \end{cases}$$

显然$cov(x_i, u_i) \neq 0$

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_k} x_{ki}$$

预测值却可能出现$\widehat{y_i} > 1$或者$\widehat{y_i} < 0$的不现实情况

两点分布（伯努利分布）

事件	1	0
概率	p	1-p

在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率

$$\begin{cases} P(y=1|\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta}) \\ P(y=0|\boldsymbol{x})=1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases}$$ 注：一般$F(\boldsymbol{x},\boldsymbol{\beta})=F(\boldsymbol{x}_i'\boldsymbol{\beta})$

$F(\boldsymbol{x},\boldsymbol{\beta})$称为连接函数($link\ function$),它将解释变量$x$和被解释变量$y$连接起来。

我们只需要保证$F(\boldsymbol{x},\boldsymbol{\beta})$是定义在$[0,1]$上的函数，就能保证$0\leq\hat{y}\leq 1$

注意：这里的定义不要理解为定义域，要理解为$F(\boldsymbol{x},\boldsymbol{\beta})$的值域是$[0,1]$

因为$E(y|\boldsymbol{x})=1\times P(y=1|\boldsymbol{x})+0\times P(y=0|\boldsymbol{x})=P(y=1|\boldsymbol{x})$

所以我们可以将$\hat{y}$可以理解为‘$y=1$’发生的概率。

连接函数的取法

$$\begin{cases} P(y=1|\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta}) \\ P(y=0|\boldsymbol{x})=1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases}$$

$F(\boldsymbol{x},\boldsymbol{\beta})$是定义在$[0,1]$上的函数

$(1)F(\boldsymbol{x},\boldsymbol{\beta})$可以取为标准正态分布的累积密度函数$(cdf)$：

$$F(\boldsymbol{x},\boldsymbol{\beta})=\Phi(\boldsymbol{x}_i'\boldsymbol{\beta})=\int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

（$probit$回归）

$(2)F(\boldsymbol{x},\boldsymbol{\beta})$可以取为$Sigmoid$函数：

$$F(\boldsymbol{x},\boldsymbol{\beta})=S(\boldsymbol{x}_i'\boldsymbol{\beta})=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

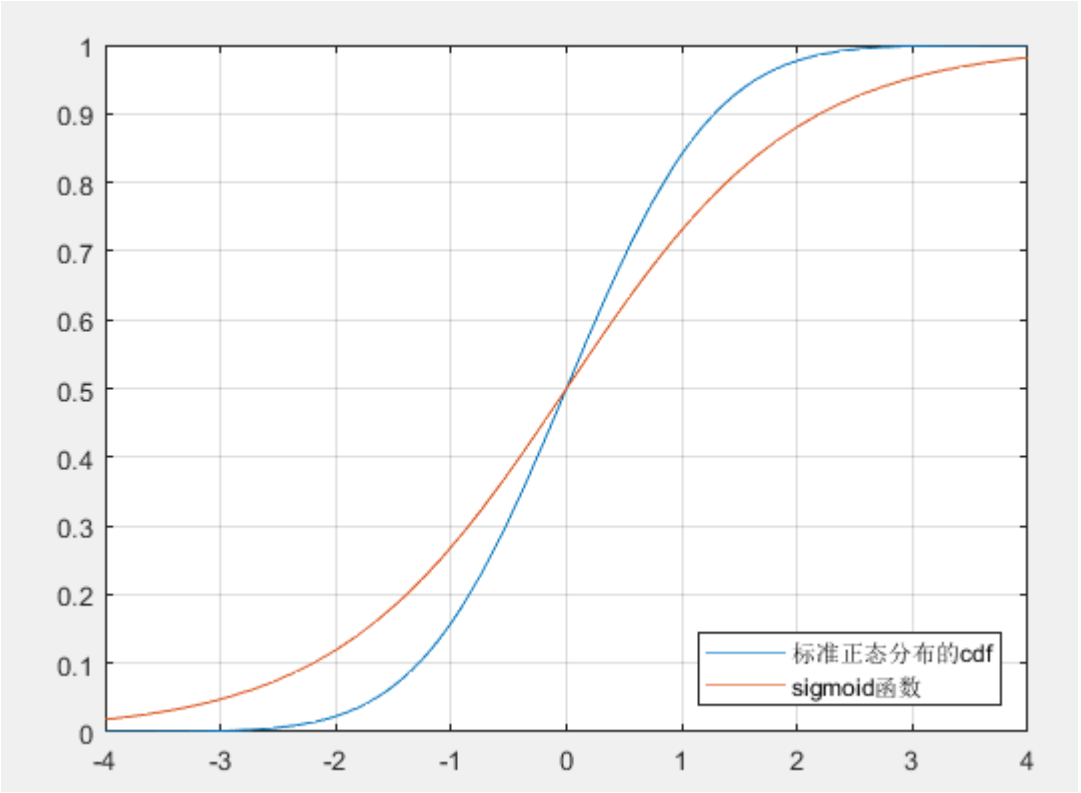
（$logistic$回归）

由于后者有解析表达式（而标准正态分布的cdf没有），所以计算logistic模型比probit模型更为方便。

函数图像对比

probit_logistic_figure.m

```
f1=@(x) normcdf(x);  % 标准正态分布的累积分布函数
fplot(f1, [-4,4]);  % 在-4到4上画出匿名函数的图形
hold on;
grid on;
f2=@(x) exp(x)/(1+exp(x));
fplot(f2, [-4,4]);
legend('标准正态分布的cdf','sigmoid函数','location','SouthEast')
```



标准正态分布的累积密度函数(cdf)：

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

$Sigmoid$ 函数：

$$S(x)=\frac{\exp(x)}{1+\exp(x)}$$

数学建模学习交流

怎么求解？

$$F(\boldsymbol{x},\boldsymbol{\beta})=S(\boldsymbol{x}_i'\boldsymbol{\beta})=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

非线性模型，使用极大似然估计方法（MLE）进行估计

$$\begin{cases} P(y=1|\boldsymbol{x})=S(\boldsymbol{x}_i'\boldsymbol{\beta}) \\ P(y=0|\boldsymbol{x})=1-S(\boldsymbol{x}_i'\boldsymbol{\beta}) \end{cases} \Rightarrow f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=\begin{cases} S(\boldsymbol{x}_i'\boldsymbol{\beta}) & ,\ y_i=1 \\ 1-S(\boldsymbol{x}_i'\boldsymbol{\beta}), & y_i=0 \end{cases}$$

写成更加紧凑的形式：

$$f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=[S(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1-S(\boldsymbol{x}_i'\boldsymbol{\beta})]^{1-y_i}$$

取对数：$\ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=y_i\ln[S(\boldsymbol{x}_i'\boldsymbol{\beta})]+(1-y_i)\ln[1-S(\boldsymbol{x}_i'\boldsymbol{\beta})]$

样本的对数似然函数：$\ln L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x})=\sum_{i=1}^{n}y_i\ln[S(\boldsymbol{x}_i'\boldsymbol{\beta})]+\sum_{i=1}^{n}(1-y_i)\ln[1-S(\boldsymbol{x}_i'\boldsymbol{\beta})]]$

可以使用数值方法（梯度下降）求解这个非线性最大化的问题。

逻辑回归的推导：https://www.bilibili.com/video/av44798895/?p=45

极大似然估计：大家可参考概率论与数理统计的教材，或搜索相应视频学习

怎么用于分类?

在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率

$$\begin{cases} P(y=1|\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta}) \\ P(y=0|\boldsymbol{x})=1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases}$$

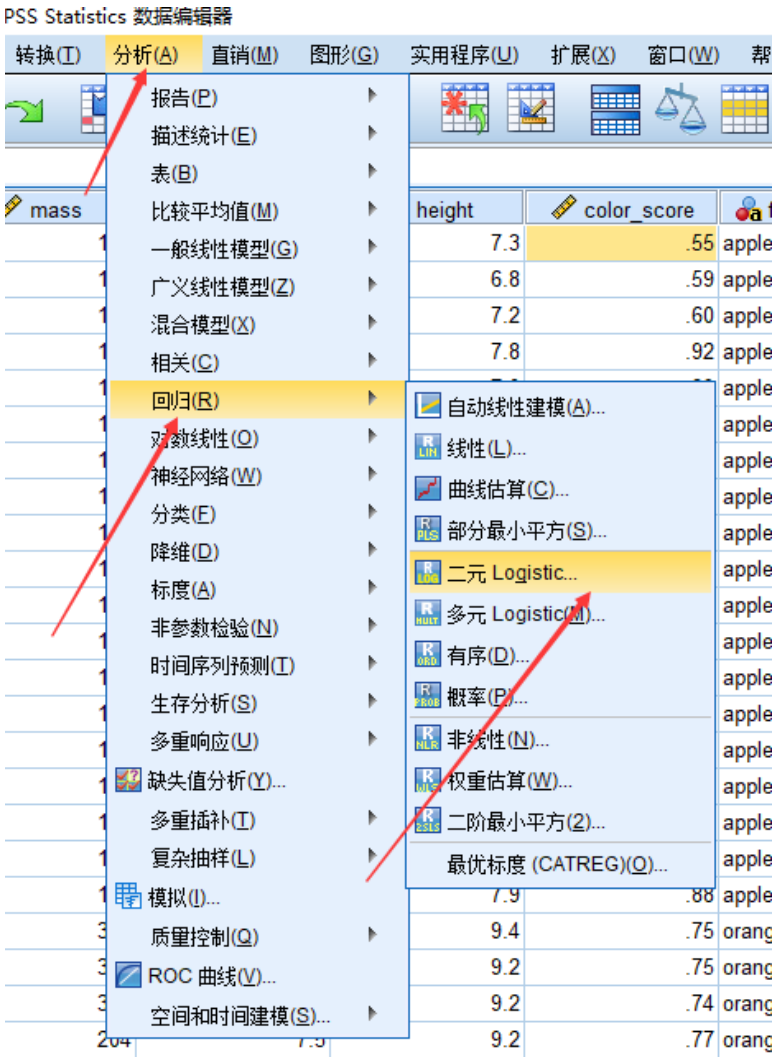
因为$E(y|\boldsymbol{x})=1\times P(y=1|\boldsymbol{x})+0\times P(y=0|\boldsymbol{x})=P(y=1|\boldsymbol{x})$

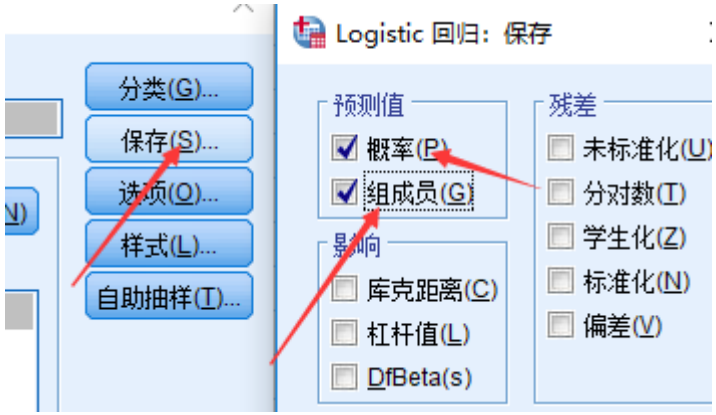
所以我们可以将$\hat{y}$可以理解为‘$y=1$’发生的概率。

$$\hat{y_i}=P(y_i=1|\boldsymbol{x})=S(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})=\frac{\exp(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})}=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ki}}}$$

如果$\hat{y_i}\geq 0.5$，则认为其预测的$y=1$;否则则认为其预测的$y=0$

Spss求解逻辑回归





预测成功率

分类表[a]

实测			预测 fruit_name=apple .00	1.00	正确百分比
步骤 1	fruit_name=apple	.00	15	4	78.9
		1.00	5	14	73.7
	总体百分比				76.3

a. 分界值为 .500

19个苹果样本中，预测出来为苹果的有14个，预测出来的正确率为73.7%;
19个橙子样本中，预测出来为橙子的有15个，预测出来的正确率为78.9%;
对于整个样本，逻辑回归的预测成功率为76.3%.

逻辑回归系数表

方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1[a]	mass	-.024	.024	.965	1	.326	.977
	width	4.307	1.844	5.452	1	.020	74.199
	height	-3.750	1.641	5.224	1	.022	.024
	color_score	9.891	5.746	2.964	1	.085	19758.273
	常量	-7.202	14.503	.247	1	.620	.001

a. 在步骤 1 输入的变量：mass, width, height, color_score。

注意:上面表格中的回归系数保留了小数点后三位，可点进去看更加精确的数据。

$$\hat{y}_i = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_k} x_{ki}}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_k} x_{ki}}}$$

$$\hat{y}_i = \frac{e^{-7.202 - 0.024 mass + 4.307 width - 3.75 height + 9.891 color\_score}}{1 + e^{-7.202 - 0.024 mass + 4.307 width - 3.75 height + 9.891 color\_score}}$$

如果$\hat{y}_i \geq 0.5$，则认为其预测的是苹果;否则则认为其预测的是橙子。

表格中新添的两列解读

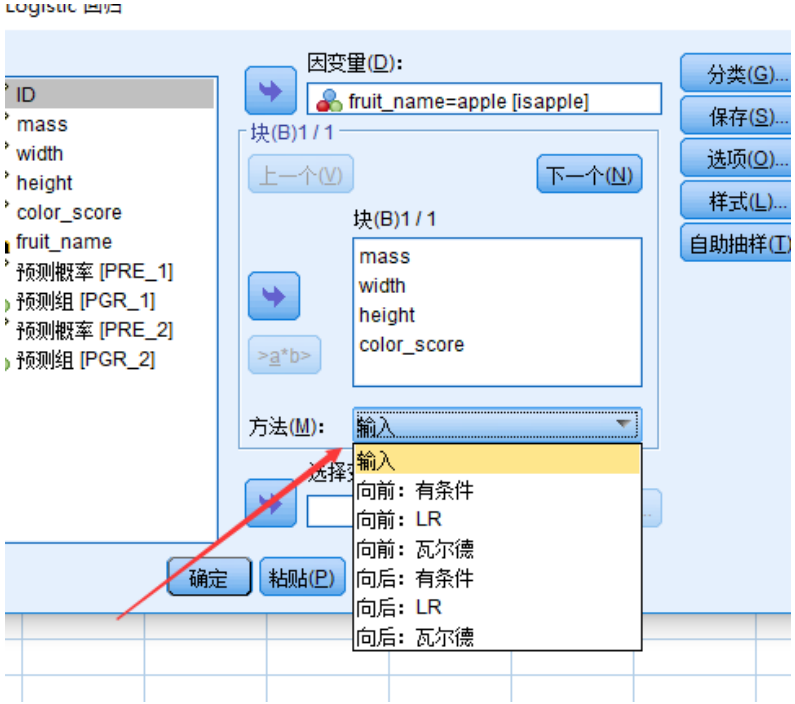
ID	mass	width	height	color_score	fruit_name	isapple	PRE_1	PGR_1	变量
1	192	8.4	7.3	.55	apple	1.00	.92283	1.00	
2	180	8.0	6.8	.59	apple	1.00	.96492	1.00	
3	176	7.4	7.2	.60	apple	1.00	.35993	.00	
4	178	7.1	7.8	.92	apple	1.00	.26898	.00	
5	172	7.4	7.0	.89	apple	1.00	.95843	1.00	
6	166	6.9	7.3	.93	apple	1.00	.59813	1.00	
7	172	7.1	7.6	.92	apple	1.00	.47319	.00	
8	154	7.0	7.1	.88	apple	1.00	.79718	1.00	
9	164	7.3	7.7	.70	apple	1.00	.16700	.00	
10	152	7.6	7.3	.69	apple	1.00	.79755	1.00	
11	156	7.7	7.1	.69	apple	1.00	.92105	1.00	
12	156	7.6	7.5	.67	apple	1.00	.58134	1.00	
13	168	7.5	7.6	.73	apple	1.00	.45786	.00	
14	162	7.5	7.1	.83	apple	1.00	.94467	1.00	
15	162	7.4	7.2	.85	apple	1.00	.90289	1.00	
16	160	7.5	7.5	.86	apple	1.00	.84316	1.00	
17	156	7.4	7.4	.84	apple	1.00	.82104	1.00	
18	140	7.3	7.1	.87	apple	1.00	.94757	1.00	

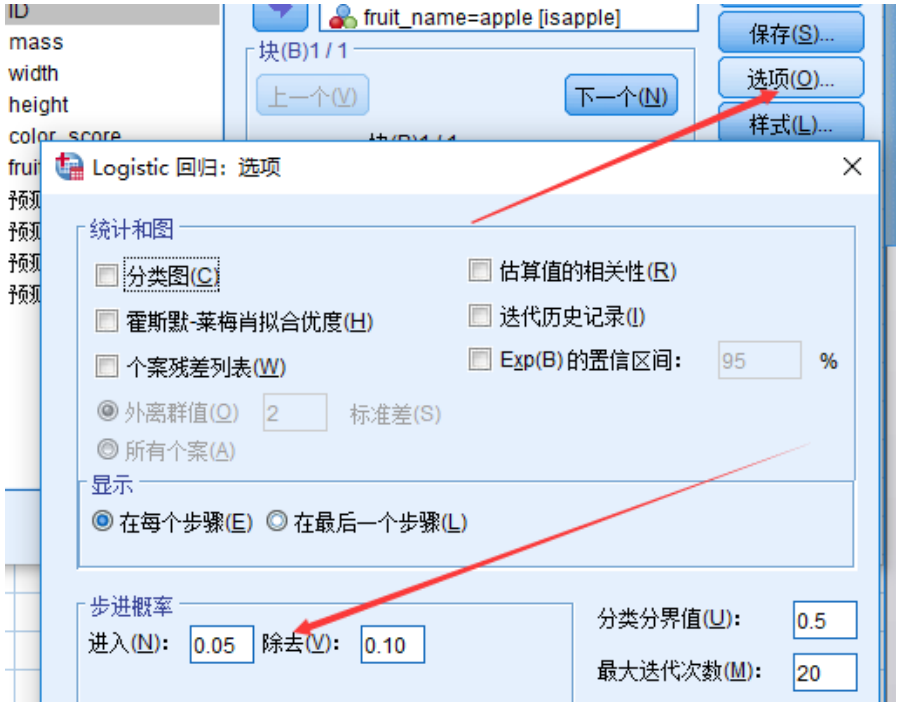
y_hat

预测的类别

$$\frac{e^{-7.201568-0.0237544\times192+4.306753\times8.4-3.749733\times7.3+9.891328\times0.55}}{1+e^{-7.201568-0.0237544\times192+4.306753\times8.4-3.749733\times7.3+9.891328\times0.55}}=0.922834$$

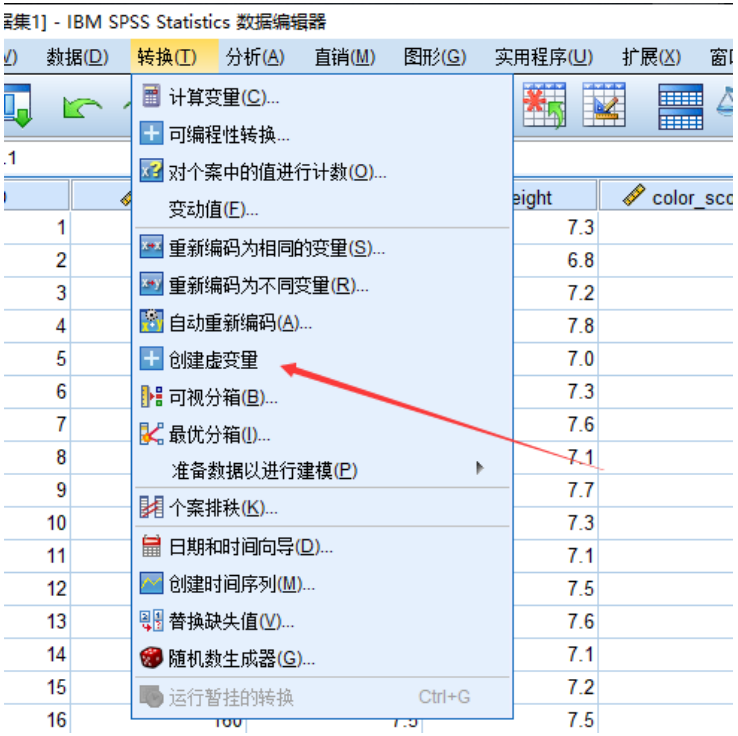
逐步回归的设置

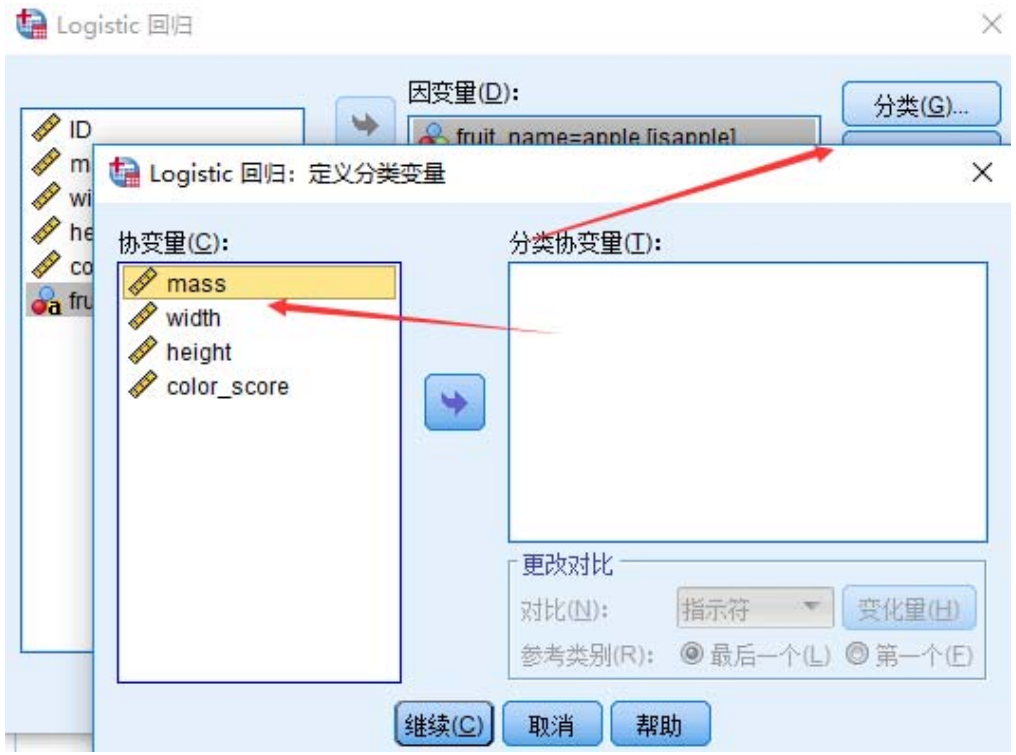




向前（向后）逐步回归可选择的统计量有所区别。
进入（或者除去）自变量的显著性水平可以自己调节。

假如自变量有分类变量怎么办?





两种方法

（1）先创建虚拟变量，然后删除任意一列以排除完全多重共线性的影响；

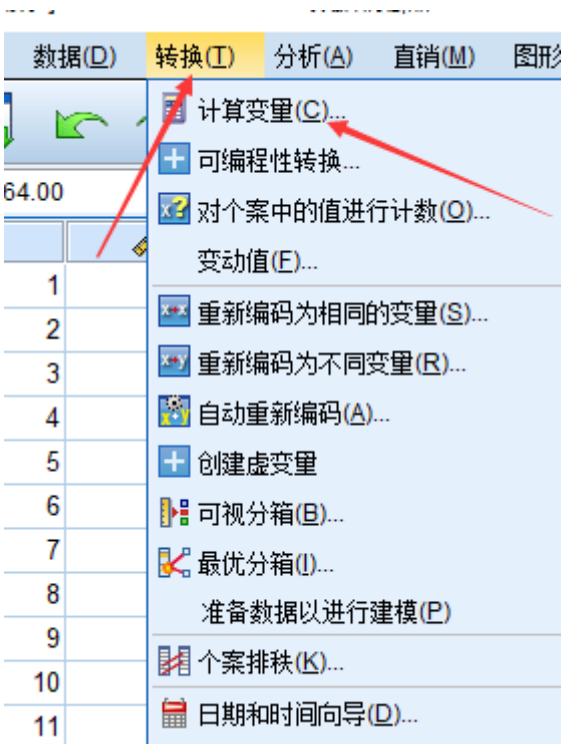
（2）直接点击分类，然后定义分类协变量，Spss会自动帮我们生成。

（如果没有生成虚拟变量这个选项，则说明SPSS没有安装到默认位置）

数学建模学习交流

预测结果较差怎么办?

可在logistic回归模型中加入平方项、交互项等。



加入了平方项后的结果

分类表[a]

实测		预测 fruit_name=apple .00	预测 fruit_name=apple 1.00	正确百分比
步骤 1	fruit_name=apple .00	19	0	100.0
	fruit_name=apple 1.00	0	19	100.0
	总体百分比			100.0

a. 分界值为 .500

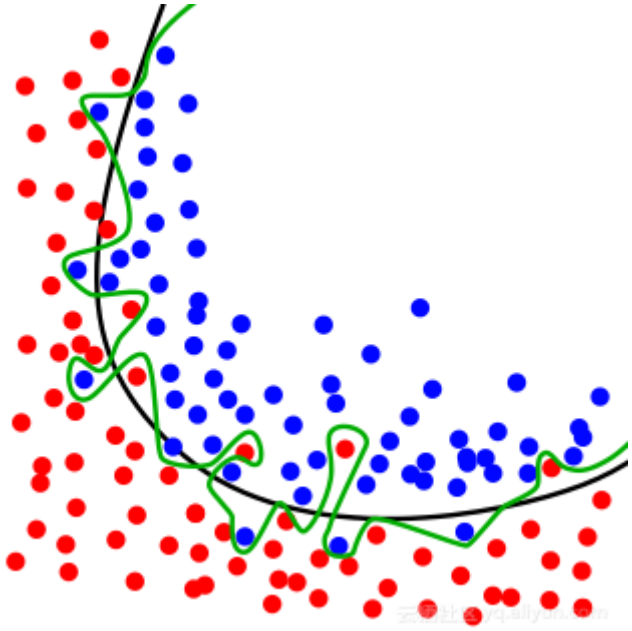
方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1[a]	mass	-11.258	4914.503	.000	1	.998	.000
	width	5272.822	3345761.455	.000	1	.999	.
	height	678.860	1139320.490	.000	1	1.000	6.683E+294
	color_score	-18786.133	3023120.896	.000	1	.995	.000
	mass2	.035	15.039	.000	1	.998	1.035
	width2	-353.067	228243.987	.000	1	.999	.000
	height2	-49.685	76875.991	.000	1	.999	.000
	color_score2	11890.132	1906867.598	.000	1	.995	.
	常量	-13667.353	8106658.615	.000	1	.999	.000

a. 在步骤 1 输入的变量：mass, width, height, color_score, mass2, width2, height2, color_score2。

PRE_1	PGR_1
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
1.00000	1.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00
.00000	.00

过拟合现象



虽然预测能力提高了，但是容易发生过拟合的现象。

对于样本数据的预测非常好，但是对于样本外的数据的预测效果可能会很差。

（是不是和龙格现象有点相似）

如何确定合适的模型

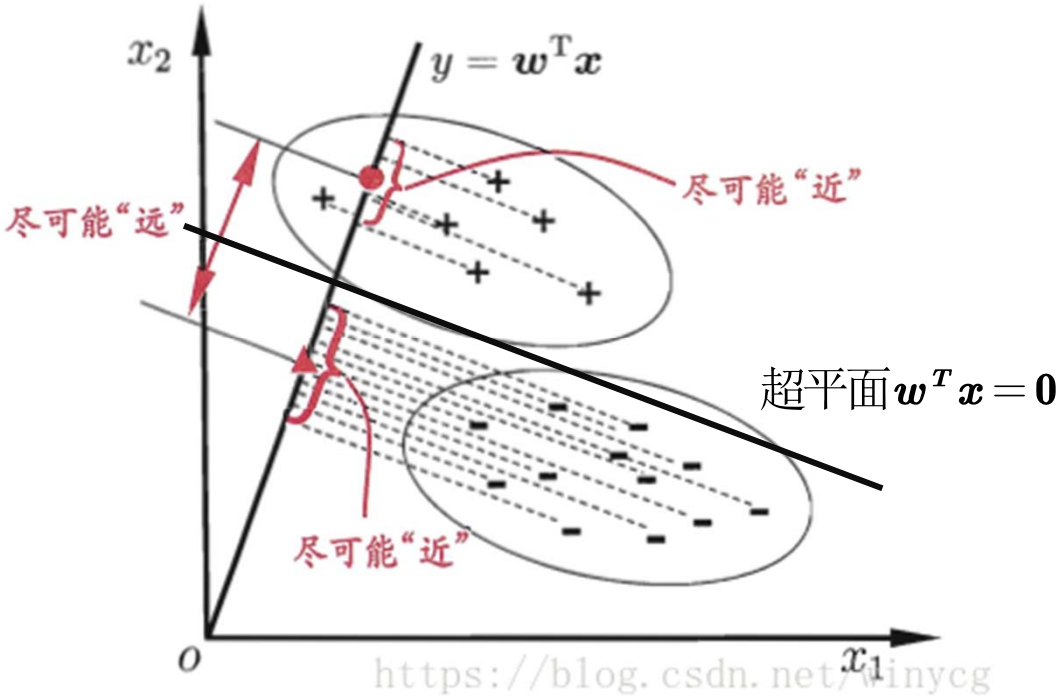
把数据分为**训练组**和**测试组**，用训练组的数据来估计出模型，再用测试组的数据来进行测试。（训练组和测试组的比例一般设置为80%和20%)

已知分类结果的水果ID为1-38，前19个为苹果，后19个为橙子。
每类水果中随机抽出3个ID作为测试组，剩下的16个ID作为训练组。
（比如：17-19、36-38这六个样本作为测试组）
比较设置不同的自变量后的模型对于测试组的预测效果。

（注意：为了消除偶然性的影响，可以对上述步骤多重复几次，最终对每个模型求一个平均的准确率，这个步骤称为交叉验证。）

Fisher线性判别分析

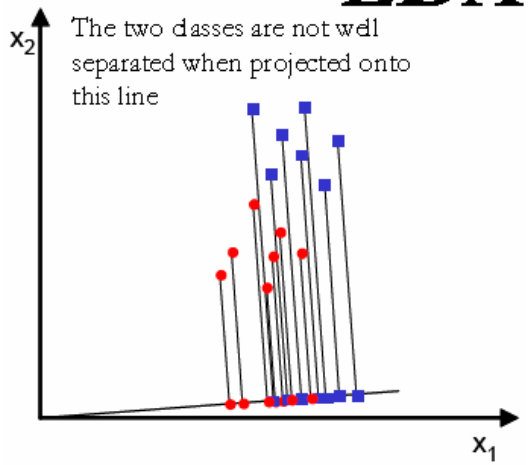
LDA(Linear Discriminant Analysis)是一种经典的线性判别方法，又称Fisher判别分析。该方法思想比较简单:给定训练集样例，设法将样例投影到一维的直线上，使得同类样例的投影点尽可能接近和密集，异类投影点尽可能远离。

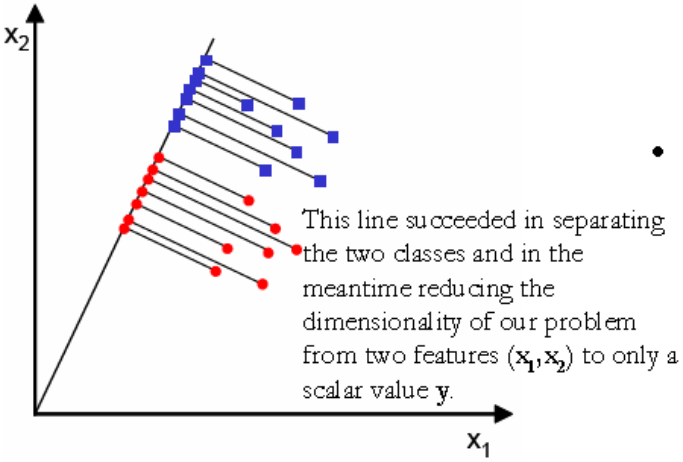


详细证明和求解步骤：https://www.bilibili.com/video/av33101528/?p=3

核心问题：找到线性系数向量ω

LDA … Two Classes





- Assume we have *m*-dimensional samples $\{x^1, x^2, \ldots, x^N\}$, N_1 of which belong to ω_1 and N_2 belong to ω_2.
- We seek to obtain a scalar **y** by projecting the samples **x** onto a line (C-1 space, C = 2).

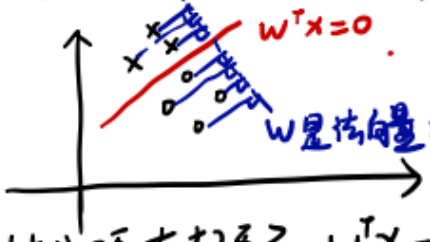
$$y = w^T x \quad where \quad x = \begin{bmatrix} x_1 \\ . \\ . \\ x_m \end{bmatrix} \quad and \quad w = \begin{bmatrix} w_1 \\ . \\ . \\ w_m \end{bmatrix}$$

- Of all the possible lines we would like to select the one that maximizes the separability of the scalars.

我自己的笔记

② Fisher 线性分类 Linear classification.

思想：



P维 → 1维

将所有的点分隔在超平面 $w^Tx=0$ 两侧 ⇔ 将每个点投影到 w这个法向量上，保证：类内小，类间大. (高内聚，松耦合).

$$X=(x_1,x_2,\cdots,x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}_{N\times p}\qquad Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_N\end{pmatrix}_{N\times 1}$$

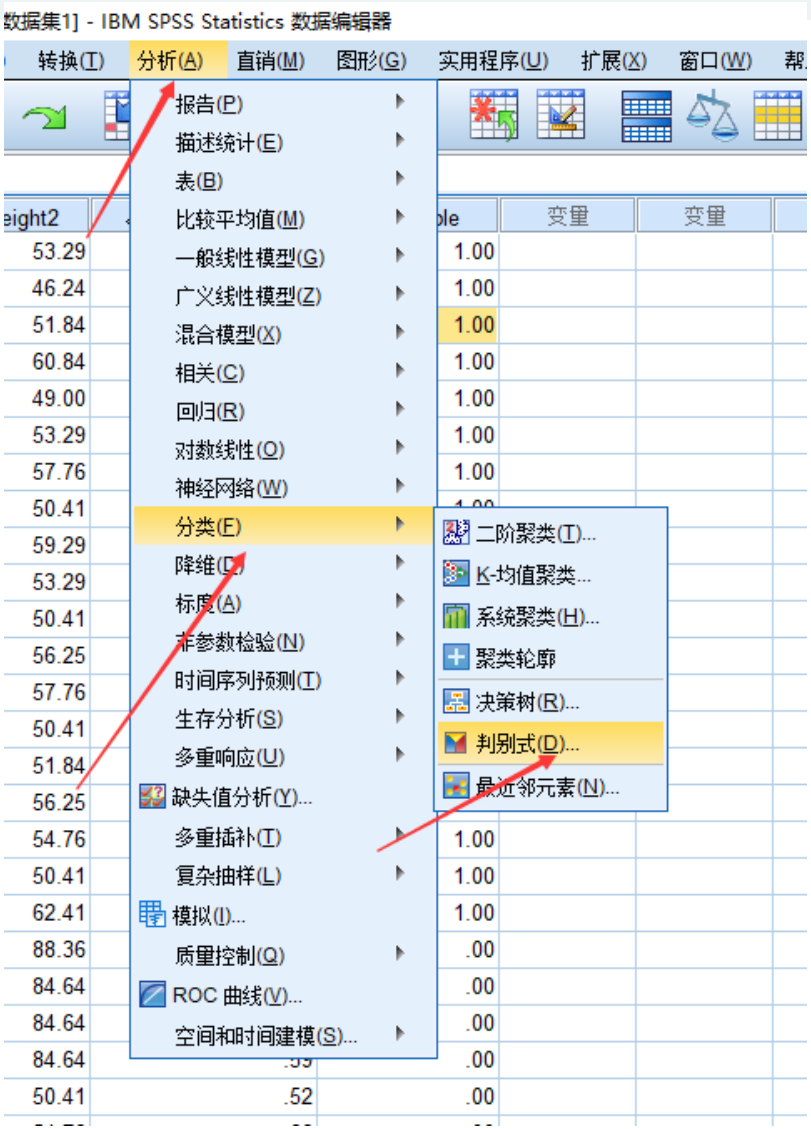
$\{(x_i,y_i)\}_{i=1}^N$, $x_i\in R^p$, $y_i\in\{\overset{C_1}{+1},\overset{C_2}{-1}\}$

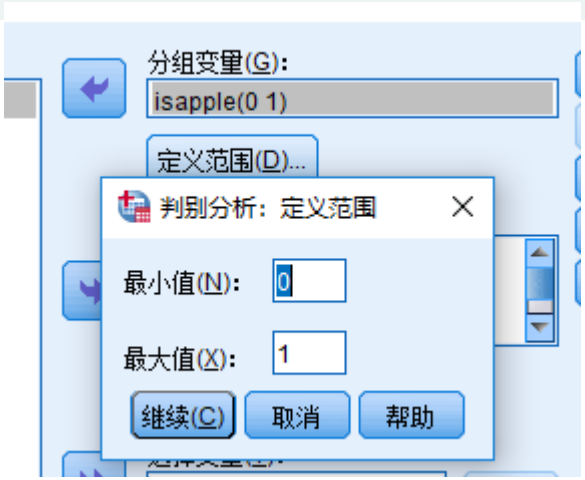
$X_{C_1}=\{x_i\mid y_i=+1\}$ $X_{C_2}=\{x_i\mid y_i=-1\}$

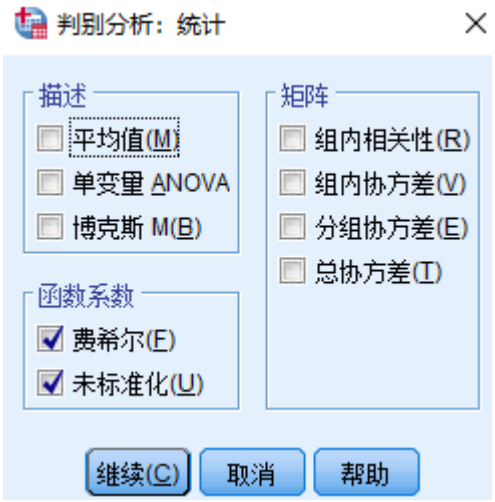
$|X_{C_1}|=N_1$, $|X_{C_2}|=N_2$ 且 $N_1+N_2=N$

本节拓展资料：白板机器学习：线性分类 .pdf

Spss操作







结果分析

典则判别函数系数

	函数 1
mass	-.021
width	3.012
height	-1.894
color_score	6.915
(常量)	-9.732

未标准化系数

线性系数向量ω

分类函数系数

	fruit_name=apple	
	.00	1.00
mass	-1.843	-1.875
width	144.004	148.579
height	48.493	45.616
color_score	310.325	320.829
(常量)	-678.418	-693.201

费希尔线性判别函数

	isapple	Dis_1	Dis1_1	Dis2_1	变量
	1.00	1.00	.09397	.90603	
	1.00	1.00	.06420	.93580	
	1.00	.00	.72767	.27233	
	1.00	.00	.68654	.31346	
	1.00	1.00	.05913	.94087	
	1.00	1.00	.44293	.55707	
	1.00	.00	.50406	.49594	
	1.00	1.00	.24571	.75429	
	1.00	.00	.80899	.19101	
	1.00	1.00	.20434	.79566	
	1.00	1.00	.09415	.90585	
	1.00	1.00	.39039	.60961	
	1.00	.00	.51349	.48651	
	1.00	1.00	.06740	.93260	
	1.00	1.00	.10985	.89015	
	1.00	1.00	.13519	.86481	
	1.00	1.00	.16740	.83260	

真实的类别

预测的类别

属于0的概率

属于1的概率

贝叶斯判别函数系数表，将样品的各参数带入2个贝叶斯判别函数，比较得出的函数值，哪个函数值较大就将该样品归于哪一类。

多分类问题

现在水果的类别一共有四种，其四个指标的平均值如下表所示：

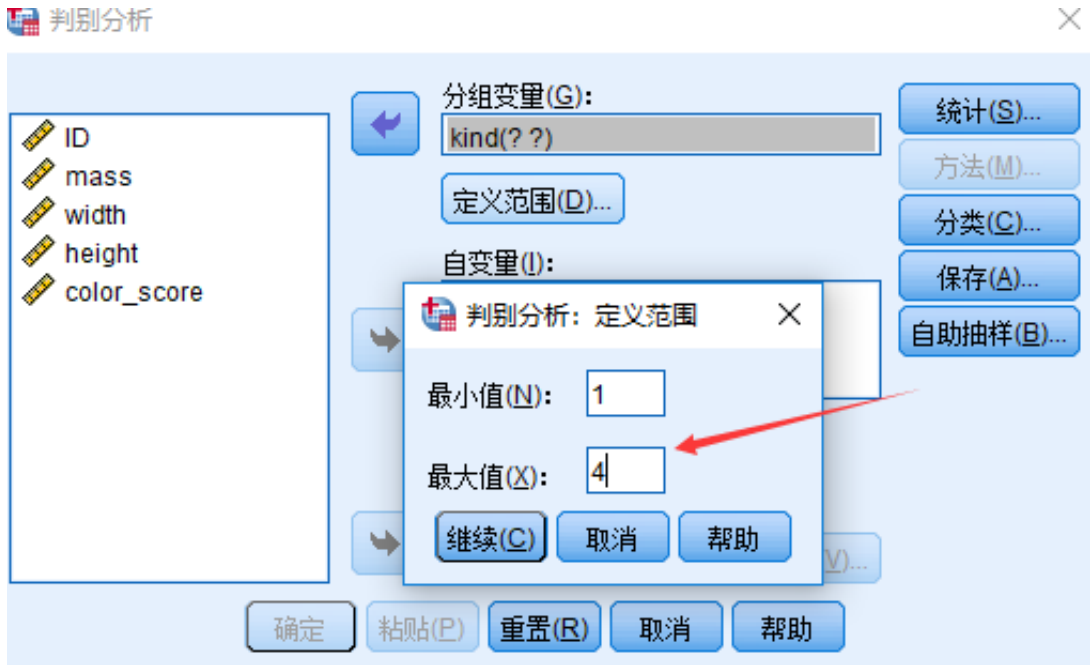
	平均值项：mass	平均值项：width	平均值项：height	平均值项：color_score	样本数
apple	165	7.46	7.34	0.78	19
lemon	150	6.51	8.86	0.72	16
mandarin（橘子）	81	5.94	4.38	0.80	5
orange	194	7.56	7.94	0.77	19

问题：对ID为60-67的八个水果进行归类。

Fisher判别分析可用于多分类

https://blog.csdn.net/z962013489/article/details/79918758

水果	Apple	Lemon	Mandarin	Orange
符号	1	2	3	4



注意：这里SPSS不能自动帮我们生成虚拟变量，我们可以在EXCEL表中使用“替换”功能来快速生成虚拟变量。

Fisher判别分析多分类的结果

kind	Dis_1	Dis1_1	Dis2_1	Dis3_1	Dis4_1
1	1	.92017	.00000	.07982	.00000
1	1	.91184	.00000	.08777	.00039
1	3	.14636	.00001	.85360	.00002
1	3	.29593	.00000	.70407	.00000
1	1	.93223	.00000	.06747	.00030
1	3	.45789	.00000	.54205	.00006
1	3	.47531	.00000	.52469	.00000
1	1	.70574	.00000	.29418	.00008
1	3	.17820	.00004	.82176	.00000
1	1	.84030	.00000	.15970	.00000
1	1	.92277	.00000	.07723	.00000
1	1	.67779	.00000	.32221	.00000
1	1	.51248	.00000	.48752	.00000
1	1	.94092	.00000	.05906	.00002
1	1	.90115	.00000	.09884	.00001
1	1	.91503	.00000	.08497	.00000
1	1	.87692	.00000	.12308	.00000
1	1	.96396	.00000	.03604	.00000
1	1	.85819	.00000	.14181	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	.99922	.00078	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	.99998	.00002	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	1.00000	.00000	.00000
2	2	.00000	.99917	.00083	.00000

分类结果[a]

			预测组成员信息				
		kind	1	2	3	4	总计
原始	计数	1	14	0	5	0	19
		2	0	16	0	0	16
		3	4	1	14	0	19
		4	0	0	0	5	5
		未分组个案	2	1	4	1	8
	%	1	73.7	.0	26.3	.0	100.0
		2	.0	100.0	.0	.0	100.0
		3	21.1	5.3	73.7	.0	100.0
		4	.0	.0	.0	100.0	100.0
		未分组个案	25.0	12.5	50.0	12.5	100.0

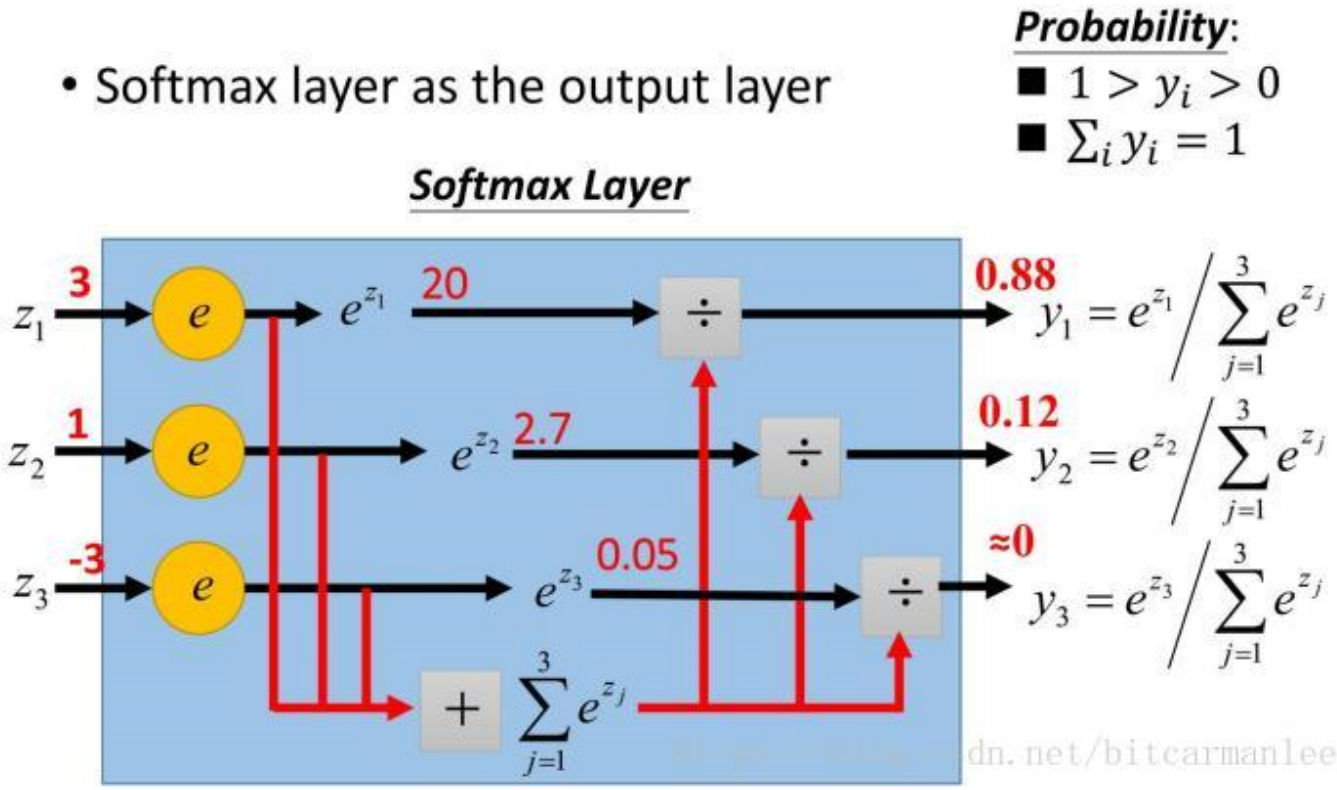
a. 正确地对 83.1% 个原始已分组个案进行了分类。

Logistic回归也可用于多分类

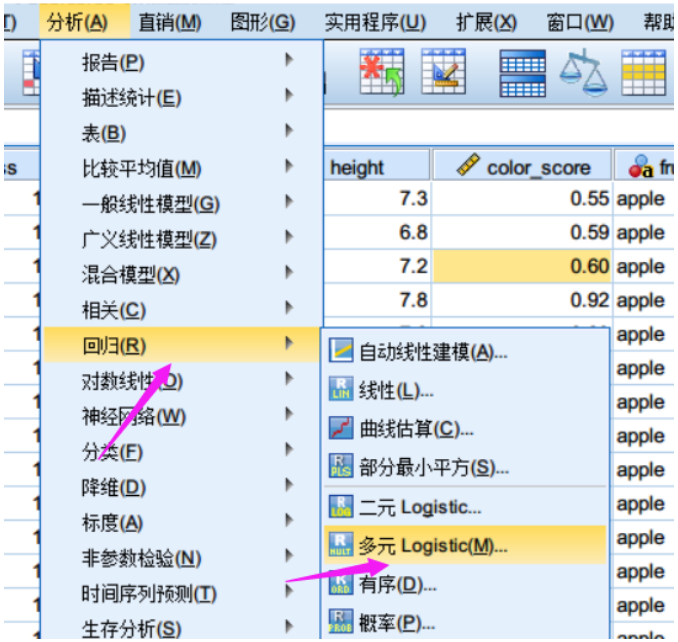
将连接函数：Sigmoid函数 推广为 Softmax函数

https://www.cnblogs.com/bonelee/p/8127411.html

https://blog.csdn.net/bitcarmanlee/article/details/82440853



Spss操作



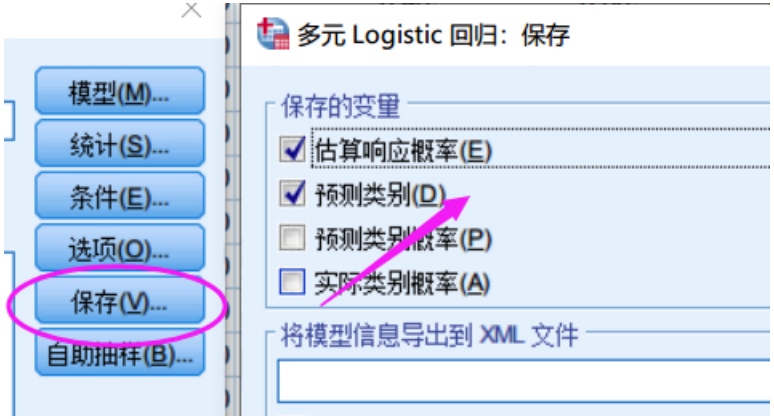
注意，这里要将几个自变量放到协变量中，视频里面的操作放到了上面的因子中是不正确的，但后续的分析思路完全相同。

补充：Spss中因子和协变量的区别

因子指分类型变量，例如性别、学历等

协变量指连续型变量，例如面积、重量等。

Spss操作



结果说明

警告

存在 177 (75.0%) 个频率为零的单元格（即，因变量级别 * 子群体）。

在海森矩阵中遇到意外的奇异性。这表明应该排除某些预测变量或者合并某些类别。

尽管发出了以上警告，但 NOMREG 过程仍将继续运行。显示的后续结果基于最后一次迭代。模型拟合度的有效性不确定。

如果遇到了左侧的警告，说明我们的数据中自变量之间存在多重共线性，或者样本中某些类别的观测值较少。
由于我们使用多元逻辑回归是出于分类的目的，因此我们重点因关注分类预测的准确度，这个警告可以忽略。

分类

实测	预测 1	2	3	4	正确百分比
1	14	0	5	0	73.7%
2	0	16	0	0	100.0%
3	4	0	15	0	78.9%
4	0	0	0	5	100.0%
总体百分比	30.5%	27.1%	33.9%	8.5%	84.7%

这张表展示了各个类别分类的准确率，总体的准确率是84.7%。
对于第2类和第4类分类效果较好，都是100%，而第1类和第3类效果较差，其中19个第1类的样本中有5个样本被误认为了第3类；而19个第3类的样本中也有4个样本被误认为了第1类.

结果说明

*多分类水果.sav [数据集1] - IBM SPSS Statistics 数据编辑器

文件(F) 编辑(E) 查看(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 扩展(X) 窗口(W) 帮助(H)

52 :

	ID	mass	width	height	color_score	fruit_name	kind	EST1_1	EST2_1	EST3_1	EST4_1	PRE_1	变量
46	46	142	7.6	7.8	0.75	orange	3	0.58	0.00	0.42	0.00	1	
47	47	150	7.1	7.9	0.75	orange	3	0.08	0.00	0.92	0.00	3	
48	48	160	7.1	7.6	0.76	orange	3	0.20	0.00	0.80	0.00	3	
49	49	154	7.3	7.3	0.79	orange	3	0.74	0.00	0.26	0.00	1	
50	50	158	7.2	7.8	0.77	orange	3	0.17	0.00	0.83	0.00	3	
51	51	144	6.8	7.4	0.75	orange	3	0.16	0.00	0.84	0.00	3	
52	52	154	7.1	7.5	0.78	orange	3	0.33	0.00	0.67	0.00	3	
53	53	180	7.6	8.2	0.79	orange	3	0.16	0.00	0.84	0.00	3	
54	54	154	7.2	7.2	0.82	orange	3	0.78	0.00	0.22	0.00	1	
55	55	86	6.2	4.7	0.80	mandarin	4	0.00	0.00	0.00	1.00	4	
56	56	84	6.0	4.6	0.79	mandarin	4	0.00	0.00	0.00	1.00	4	
57	57	80	5.8	4.3	0.77	mandarin	4	0.00	0.00	0.00	1.00	4	
58	58	80	5.9	4.3	0.81	mandarin	4	0.00	0.00	0.00	1.00	4	
60	60	158	7.1	7.6	0.72		.	0.15	0.00	0.85	0.00	3	
61	61	190	7.5	7.9	0.77		.	0.19	0.00	0.81	0.00	3	
62	62	189	7.6	7.7	0.77		.	0.45	0.00	0.55	0.00	3	
63	63	160	7.9	6.9	0.65		.	0.97	0.00	0.03	0.00	1	
64	64	150	6.4	8.6	0.72		.	0.00	1.00	0.00	0.00	2	
65	65	87	6.9	4.8	0.82		.	0.00	0.00	0.00	1.00	4	
66	66	92	7.1	5.4	0.79		.	0.19	0.00	0.00	0.81	4	
67	67	120	6.8	7.8	0.68		.	0.00	1.00	0.00	0.00	2	
68													
70													
71													

预测的属于每一个类别的概率

返回到我们的数据列表，可以看出Spss给我们输出了属于每一类的概率，并将概率最大的那个类别作为我们的预测结果。

课后作业

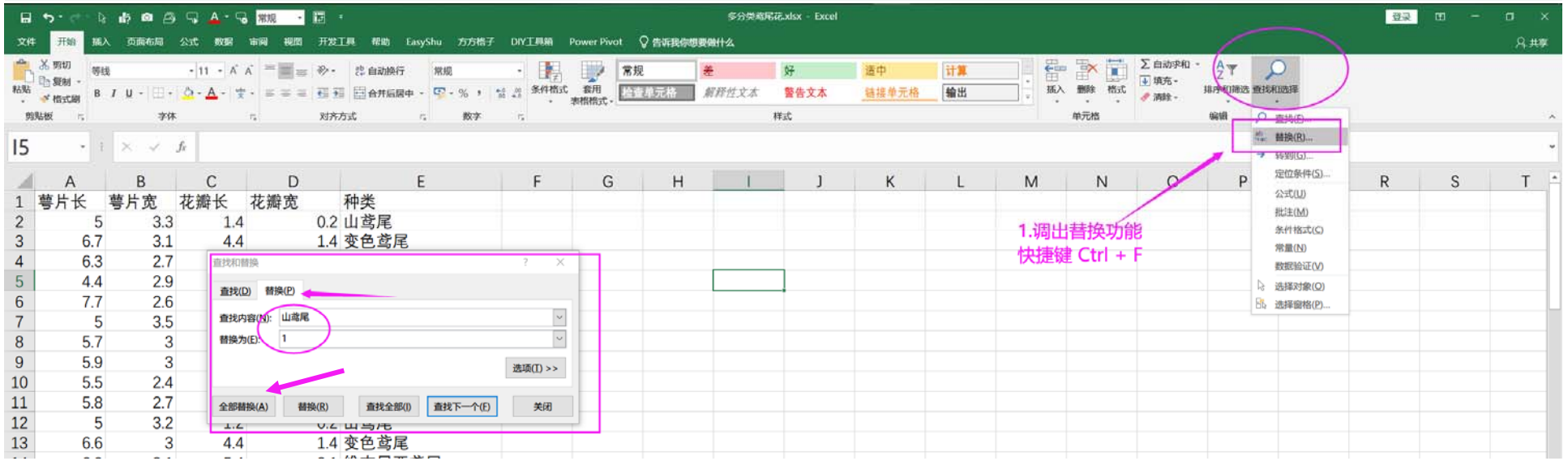
鸢尾花有很多种分类，它们一般通过花萼长度、花萼宽度、花瓣长度、花瓣宽度进行区分。

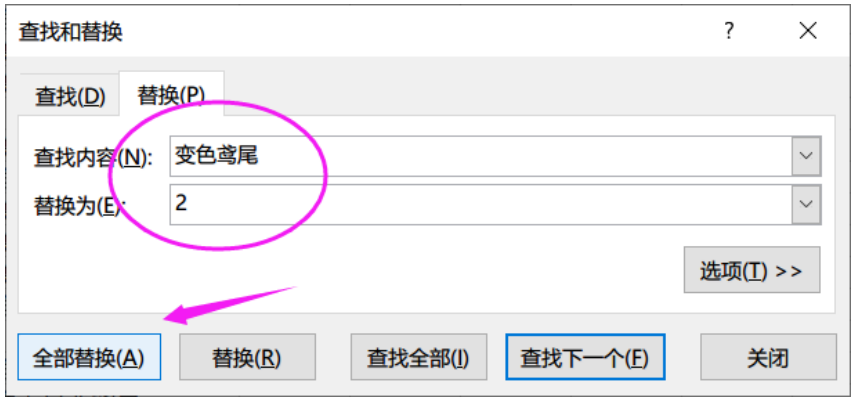
请根据数据建立合适的模型完成鸢尾花的分类，并预测剩下六朵花对应的种类。

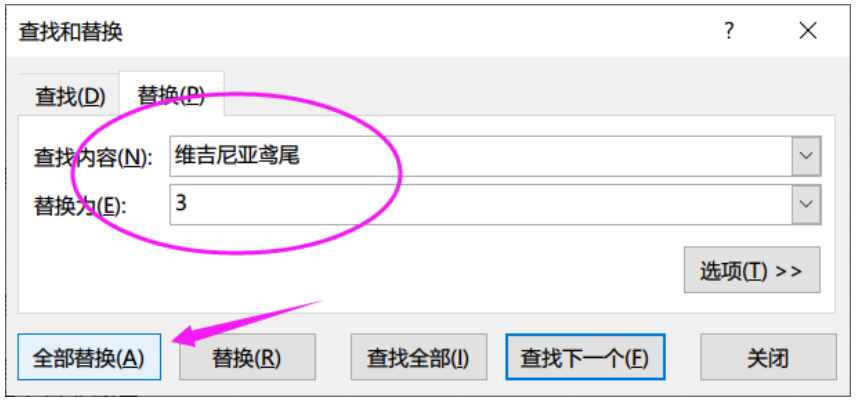
4.9	3.1	1.5	0.1	山鸢尾
6.3	2.9	5.6	1.8	维吉尼亚鸢尾
5.4	3.7	1.5	0.2	山鸢尾
6.3	2.3	4.4	1.3	变色鸢尾
6.4	2.8	5.6	2.1	维吉尼亚鸢尾
5.2	3.4	1.4	0.2	山鸢尾

注：表中空着的六朵花真实的分类如上表所示。

参考答案





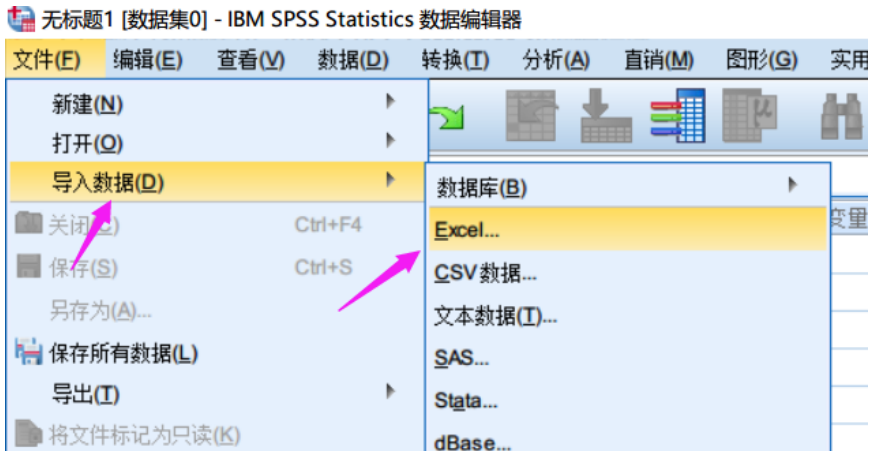


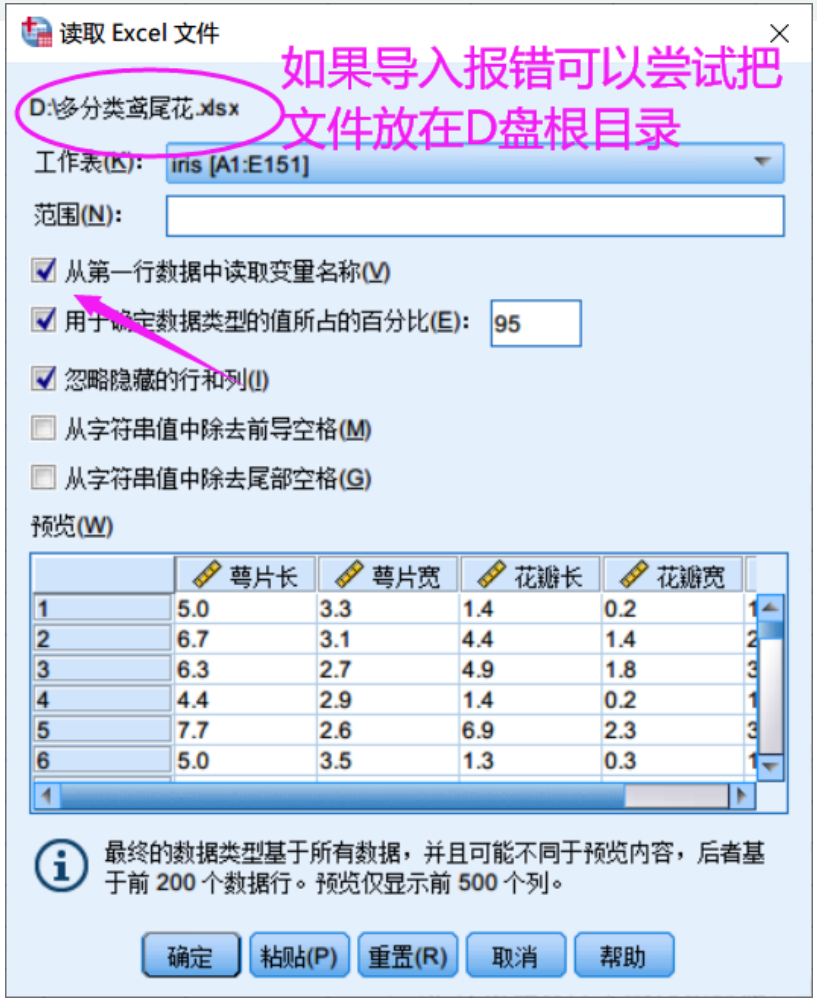
参考答案

替换后的数据，接下来就可以导入到excel了。

	A	B	C	D	E
1	萼片长	萼片宽	花瓣长	花瓣宽	种类
2	5	3.3	1.4	0.2	1
3	6.7	3.1	4.4	1.4	2
4	6.3	2.7	4.9	1.8	3
5	4.4	2.9	1.4	0.2	1
6	7.7	2.6	6.9	2.3	3
7	5	3.5	1.3	0.3	1
8	5.7	3	4.2	1.2	2
9	5.9	3	5.1	1.8	3
10	5.5	2.4	3.8	1.1	2
11	5.8	2.7	3.9	1.2	2
12	5	3.2	1.2	0.2	1
13	6.6	3	4.4	1.4	2
14	6.9	3.1	5.4	2.1	3
15	5.6	3	4.1	1.3	2
16	5.1	2.5	3	1.1	2
17	6.1	2.8	4	1.3	2
18	6.9	3.1	5.1	2.3	3
19	5	3.6	1.4	0.2	1
20	6.7	3	5.2	2.3	3
21	4.9	2.5	4.5	1.7	3

参考答案





另外，如果导入excel数据文件比较卡的话，可以先将数据另存为csv文件，然后再使用SPSS的导入csv数据的功能。

参考答案

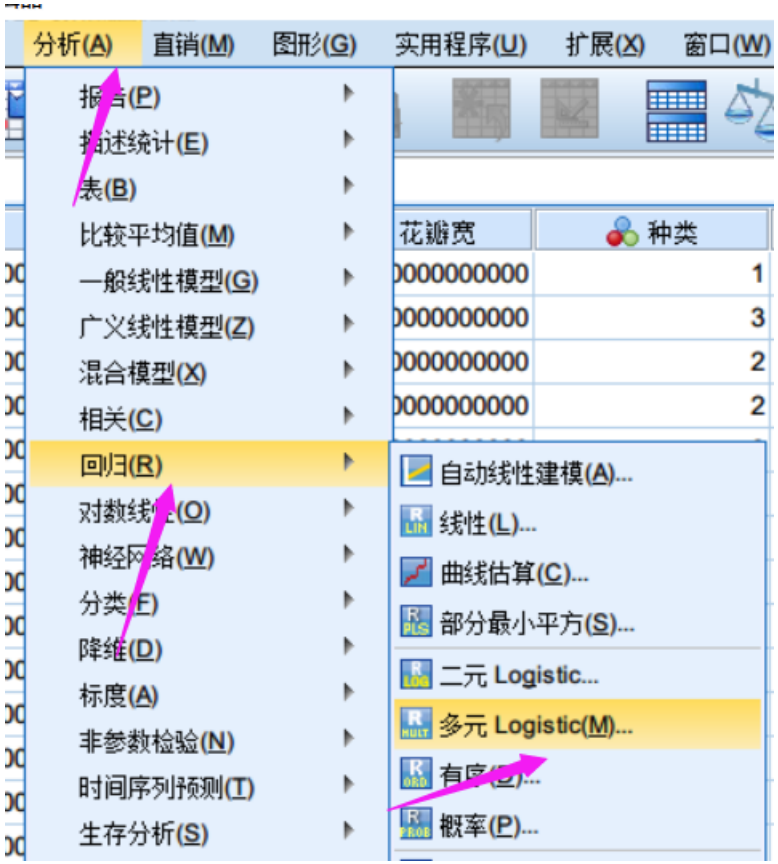
导入进来的数据

	萼片长	萼片宽	花瓣长	花瓣宽	种类	变量
1	5.00000000000000	3.30000000000000	1.40000000000000	0.20000000000000	1	
2	6.70000000000000	3.10000000000000	4.40000000000000	1.40000000000000	2	
3	6.30000000000000	2.70000000000000	4.90000000000000	1.80000000000000	3	
4	4.40000000000000	2.90000000000000	1.40000000000000	0.20000000000000	1	
5	7.70000000000000	2.60000000000000	6.90000000000000	2.30000000000000	3	
6	5.00000000000000	3.50000000000000	1.30000000000000	0.30000000000000	1	
7	5.70000000000000	3.00000000000000	4.20000000000000	1.20000000000000	2	
8	5.90000000000000	3.00000000000000	5.10000000000000	1.80000000000000	3	
9	5.50000000000000	2.40000000000000	3.80000000000000	1.10000000000000	2	
10	5.80000000000000	2.70000000000000	3.90000000000000	1.20000000000000	2	
11	5.00000000000000	3.20000000000000	1.20000000000000	0.20000000000000	1	
12	6.60000000000000	3.00000000000000	4.40000000000000	1.40000000000000	2	
13	6.90000000000000	3.10000000000000	5.40000000000000	2.10000000000000	3	
14	5.60000000000000	3.00000000000000	4.10000000000000	1.30000000000000	2	
15	5.10000000000000	2.50000000000000	3.00000000000000	1.10000000000000	2	
16	6.10000000000000	2.80000000000000	4.00000000000000	1.30000000000000	2	
17	6.90000000000000	3.10000000000000	5.10000000000000	2.30000000000000	3	
18	5.00000000000000	3.60000000000000	1.40000000000000	0.20000000000000	1	
19	6.70000000000000	3.00000000000000	5.20000000000000	2.30000000000000	3	
20	4.90000000000000	2.50000000000000	4.50000000000000	1.70000000000000	3	
21	6.80000000000000	3.20000000000000	5.90000000000000	2.30000000000000	3	
22	6.50000000000000	3.00000000000000	5.50000000000000	1.80000000000000	3	
23	4.80000000000000	3.40000000000000	1.90000000000000	0.20000000000000	1	
24	4.60000000000000	3.40000000000000	1.40000000000000	0.30000000000000	1	
25	7.00000000000000	3.20000000000000	4.70000000000000	1.40000000000000	2	

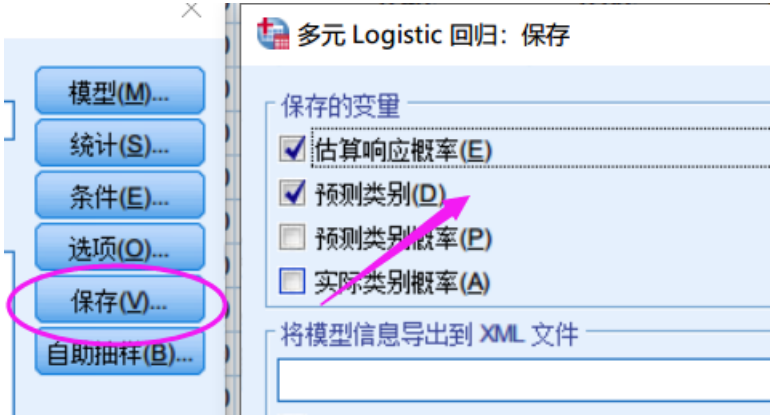
137	6.00000000000000	3.40000000000000	4.50000000000000	1.60000000000000	2
138	5.80000000000000	2.80000000000000	5.10000000000000	2.40000000000000	3
139	4.70000000000000	3.20000000000000	1.60000000000000	0.20000000000000	1
140	5.90000000000000	3.00000000000000	4.20000000000000	1.50000000000000	2
141	6.10000000000000	2.80000000000000	4.70000000000000	1.20000000000000	2
142	6.10000000000000	2.60000000000000	5.60000000000000	1.40000000000000	3
143	5.00000000000000	3.00000000000000	1.60000000000000	0.20000000000000	1
144	7.70000000000000	2.80000000000000	6.70000000000000	2.00000000000000	3
145	4.90000000000000	3.10000000000000	1.50000000000000	0.10000000000000	.
146	6.30000000000000	2.90000000000000	5.60000000000000	1.80000000000000	.
147	5.40000000000000	3.70000000000000	1.50000000000000	0.20000000000000	.
148	6.30000000000000	2.30000000000000	4.40000000000000	1.30000000000000	.
149	6.40000000000000	2.80000000000000	5.60000000000000	2.10000000000000	.
150	5.20000000000000	3.40000000000000	1.40000000000000	0.20000000000000	.
151					
152					

这是我们要预测的

参考答案：以多元逻辑回归为例



参考答案



参考答案

警告

数据中可能存在准完全分隔。最大似然估算值不存在，或者某些参数估算值无穷大。

尽管发出了以上警告，但 NOMREG 过程仍将继续运行。显示的后续结果基于最后一次迭代。模型拟合度的有效性不确定。

左侧这个警告说明我们的数据区分度非常高，存在准完全分隔说明样本划分的比较彻底，这一般对于分类结果而言是好事情。

分类

	预测			
实测	1	2	3	正确百分比
1	47	0	0	100.0%
2	0	48	1	98.0%
3	0	1	47	97.9%
总体百分比	32.6%	34.0%	33.3%	98.6%

这张表展示了各个类别分类的准确率，总体的准确率是98.6%。
对于第1类分类效果是100%，而第2类和第3类都只判断错了1个，效果也是非常好的.

参考答案

142	6.10000000000000	2.60000000000000	5.60000000000000	1.40000000000000	3	0.00	0.03	0.97	3
143	5.00000000000000	3.00000000000000	1.60000000000000	0.20000000000000	1	1.00	0.00	0.00	1
144	7.70000000000000	2.80000000000000	6.70000000000000	2.00000000000000	3	0.00	0.00	1.00	3
145	4.90000000000000	3.10000000000000	1.50000000000000	0.10000000000000	.	1.00	0.00	0.00	1
146	6.30000000000000	2.90000000000000	5.60000000000000	1.80000000000000	.	0.00	0.00	1.00	3
147	5.40000000000000	3.70000000000000	1.50000000000000	0.20000000000000	.	1.00	0.00	0.00	1
148	6.30000000000000	2.30000000000000	4.40000000000000	1.30000000000000	.	0.00	1.00	0.00	2
149	6.40000000000000	2.80000000000000	5.60000000000000	2.10000000000000	.	0.00	0.00	1.00	3
150	5.20000000000000	3.40000000000000	1.40000000000000	0.20000000000000	.	1.00	0.00	0.00	1
151									
152									

这6个未知类别的预测结果也出来了，分别是1 3 1 2 3 1，而在最开始替换时，1对应山鸢尾，2对应变色鸢尾，3对应维吉尼亚鸢尾 。

另外大家也可以使用Fisher判别分析进行多分类，这里就不再赘述了。

课后思考

最近很火的垃圾分类问题，我们能否设计一个分类模型出来？

提示：有兴趣的同学可以搜索深度学习进行垃圾分类，可能会出现在计算机视觉、数据挖掘或者大数据分析的竞赛中。